Penn Engineering

DEPARTMENT OF THE NAVY USN ONR Science & Technology

PreSTA: Preventing Malicious Behavior Using Spatio-Temporal Reputation

Andrew G. West

November 4, 2009

ONR-MURI Presentation

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE
04 NOV 2009

2. REPORT TYPE

3. DATES COVERED
00-00-2009 to 00-00-2009

4. TITLE AND SUBTITLE
PreSTA: Preventing Malicious Behavior Using Spatio-Temporal Reputation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pennsylvania,School of Engineering and Applied Science,220 South 33rd Street ,Philadelphia,PA,19104-6391

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
ONR-MURI Presentation, Nov 2009

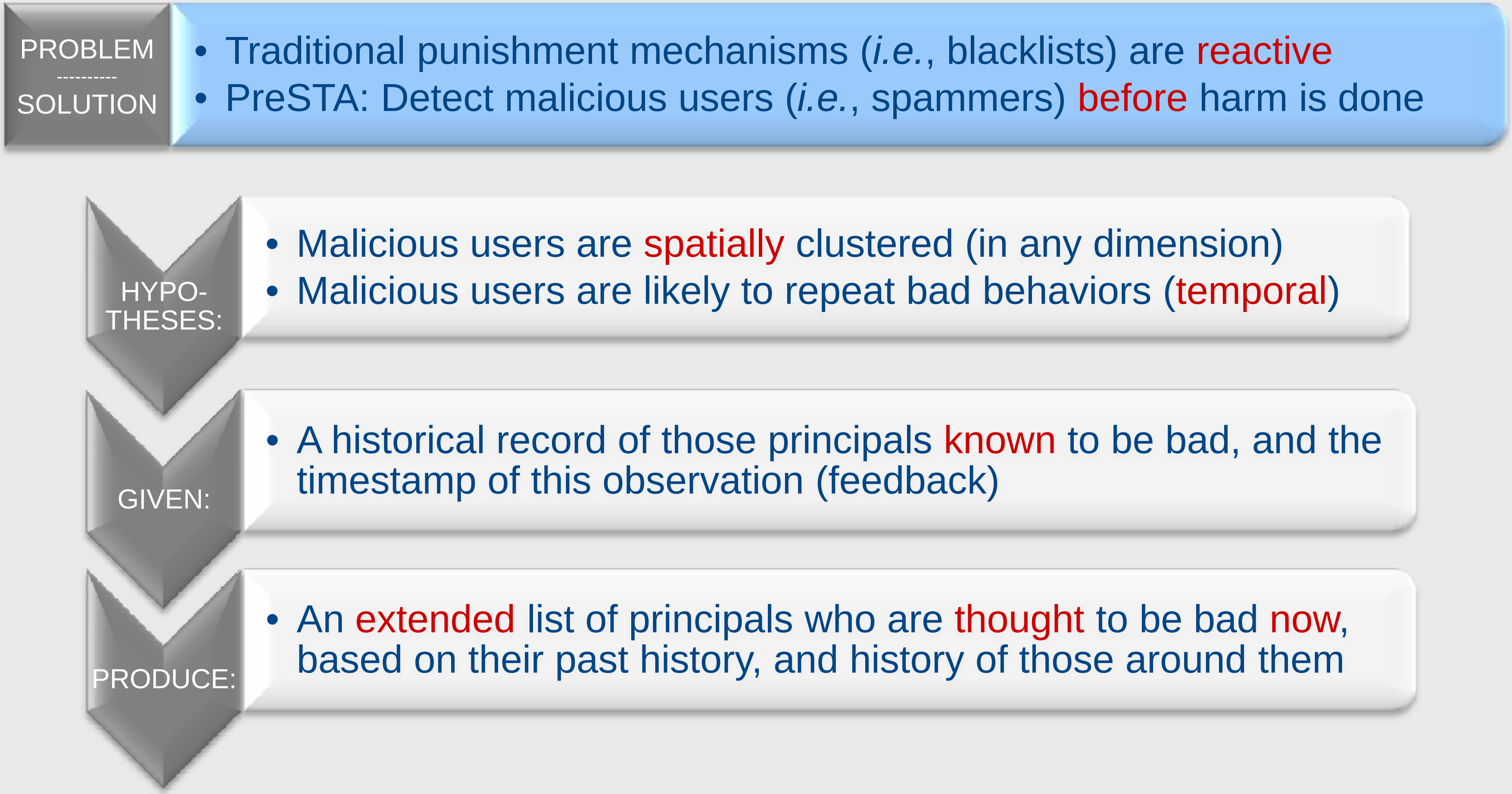
14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	17. LIMITATION OF ABSTRACT **Same as Report (SAR)**	18. NUMBER OF PAGES **23**	19a. NAME OF RESPONSIBLE PERSON

PreSTA: Preventative Spatio-Temporal Aggregation



PROBLEM

SOLUTION

- Traditional punishment mechanisms (*i.e.*, blacklists) are reactive
- PreSTA: Detect malicious users (*i.e.*, spammers) before harm is done

HYPO-THESES:

- Malicious users are spatially clustered (in any dimension)
- Malicious users are likely to repeat bad behaviors (temporal)

GIVEN:

- A historical record of those principals known to be bad, and the timestamp of this observation (feedback)

PRODUCE:

- An extended list of principals who are thought to be bad now, based on their past history, and history of those around them

TALK OUTLINE

PreSTA Running Example: Spam Detection

- Spatio-temporal properties of spam mail
- Basis for spatial groupings
- Calculating and combining reputations
- Classifier performance

Generalizing PreSTA: Additional Use-Cases for Model

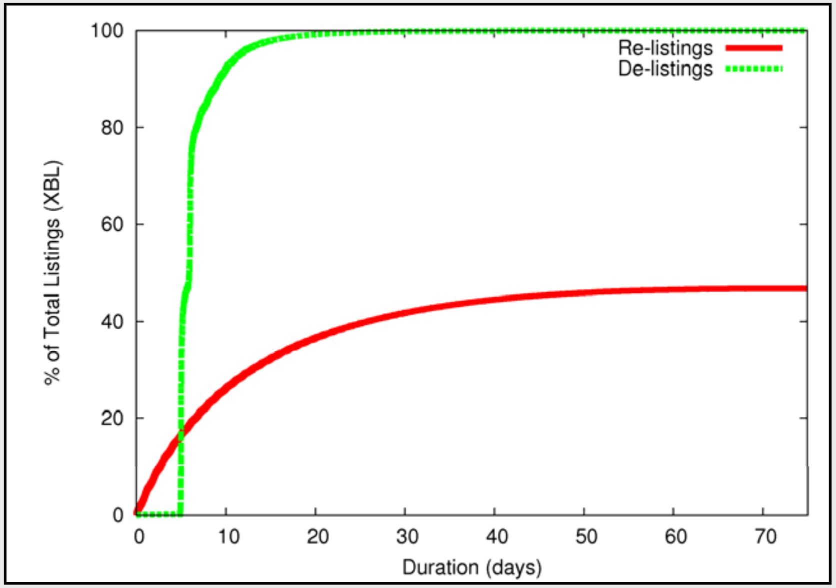
- Malicious editors on Wikipedia
- Applicability to the QuanTM model
- General PreSTA use-case criteria

Conclusions & References

SPAM: TEMPORAL PROPERTIES

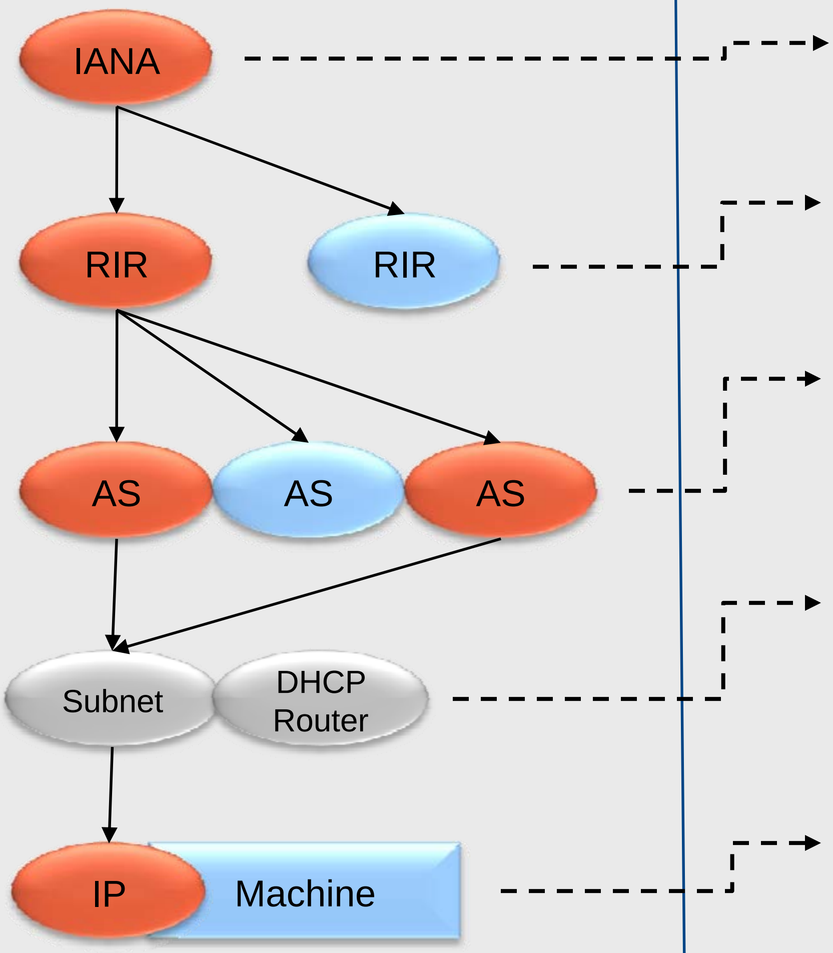
TEMPORAL: Bad Guys Repeat Bad Behaviors

- Spammers want to maximize utilization of available IP addresses, leading to re-use
- Bot-nets will compromise a machine until patched
- Blacklist entries have predictable duration (~6 days), making for trivial recycling



- Most mail servers have static IP addresses, so IP acts as a persistent identifier – though we later discuss DHCP considerations

IP DELEGATION HIERARCHY



(1) Internet Assigned Numbers Auth.: Controls all IP delegation (root of trust)

(2) Regional Internet Registries: Continent-level equivalent of the IANA

(3) Autonomous Systems (ISPs): Broadcast the IPs they control via the Border Gateway Protocol (BGP)

(4) Local routers distribute addresses from some pool (*i.e.*, a /24). Such subnet boundaries are NOT known

(5) Individual IP: Over time a single IP may have multiple inhabitants (due to dynamic nature – DHCP)

SPATIAL GROUPINGS

~~IANA /RIR~~

- The IANA and RIR granularity are too broad to be of relevant use

AS

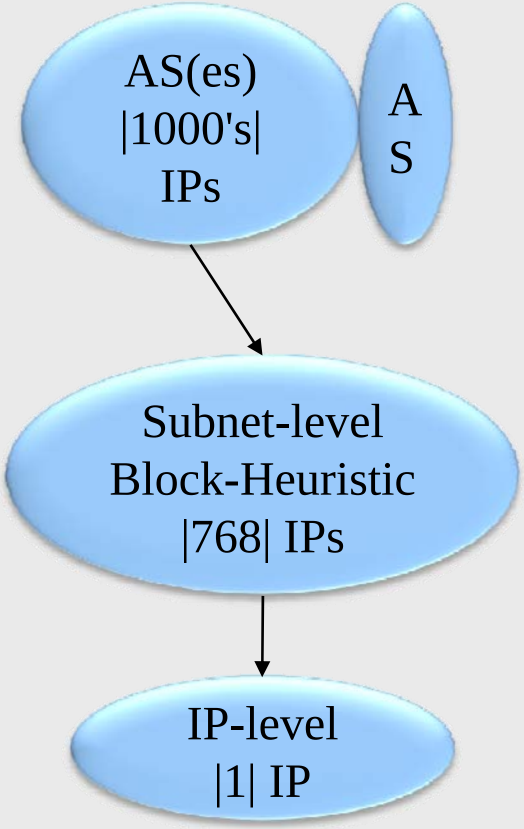
- What AS(es) are broadcasting IP?
- An IP may have 0, 1, or 2+ homes

BLOCK

- What is /24 (256 IP) membership?
- Valuate that block and two adjacent
- Estimation of subnet membership

IP

- Simplest case. Little spatial value.
- Due to DHCP, may have multiple inhabitants over time, though



AS(es) |1000's| IPs

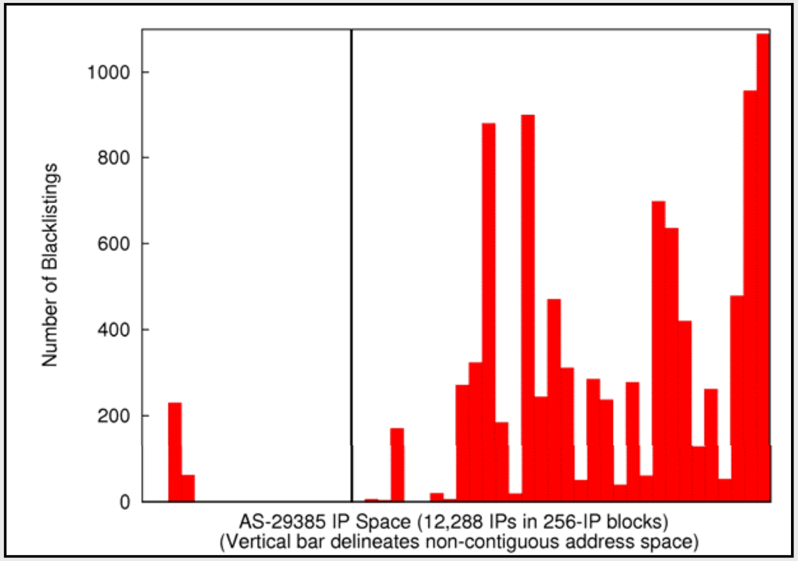
AS

Subnet-level Block-Heuristic |768| IPs

IP-level |1| IP

SPAM: SPATIAL PROPERTIES

SPATIAL: Bad Guys Live in Close Proximity [3] (IP)

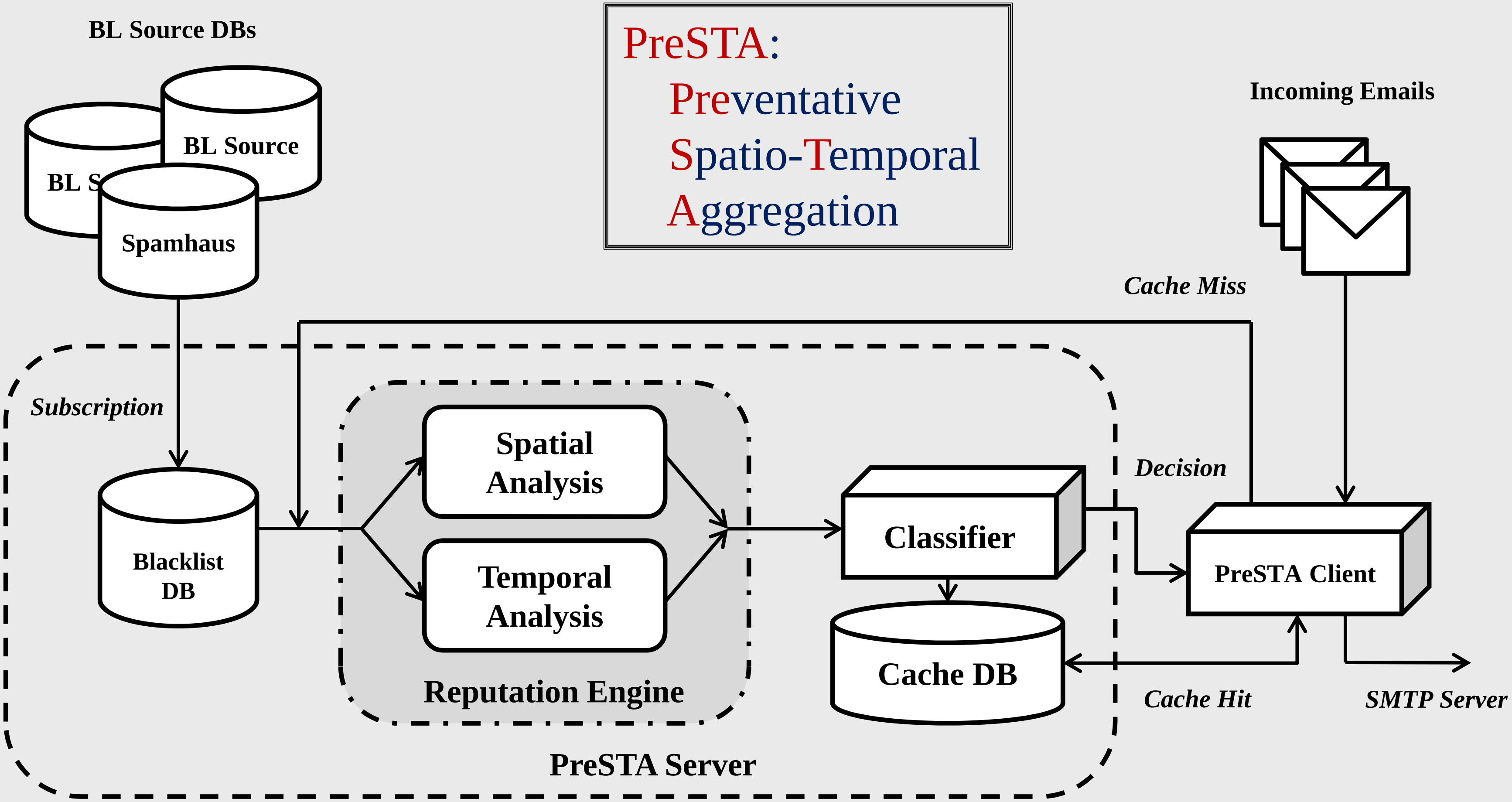


- Some ISPs/AS willing to trade behavioral leniency for compensation: McColo Corp. and 3FN
- Some geographical jurisdictions are more lenient than others (and this maps into IP space)

- As IPs become BL'ed, operations must shift to 'fresh' addresses, likely those from the same allocation (*i.e.*, subnets)

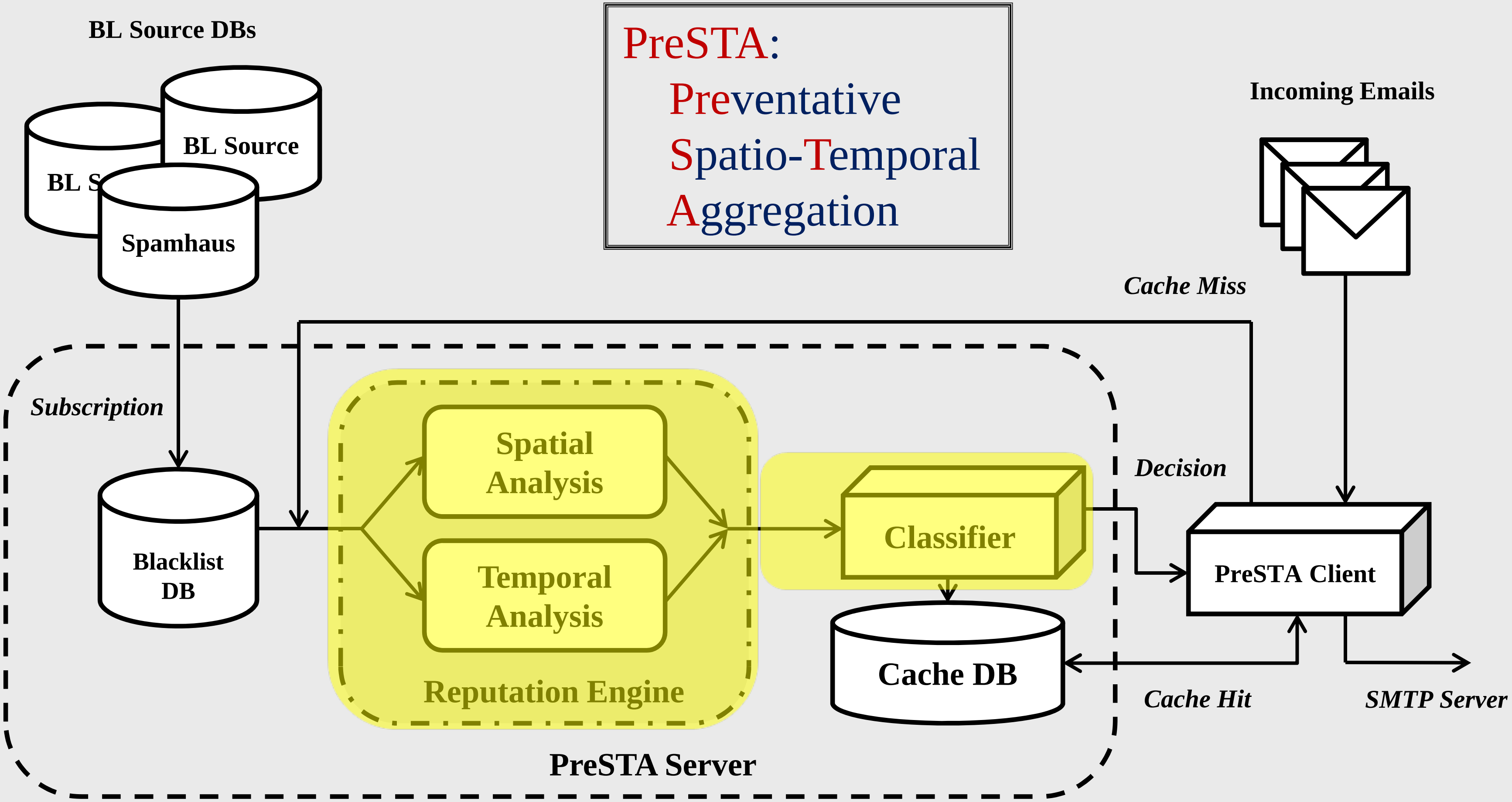
PreSTA: SPAM USAGE

PreSTA:
Preventative
Spatio-Temporal
Aggregation

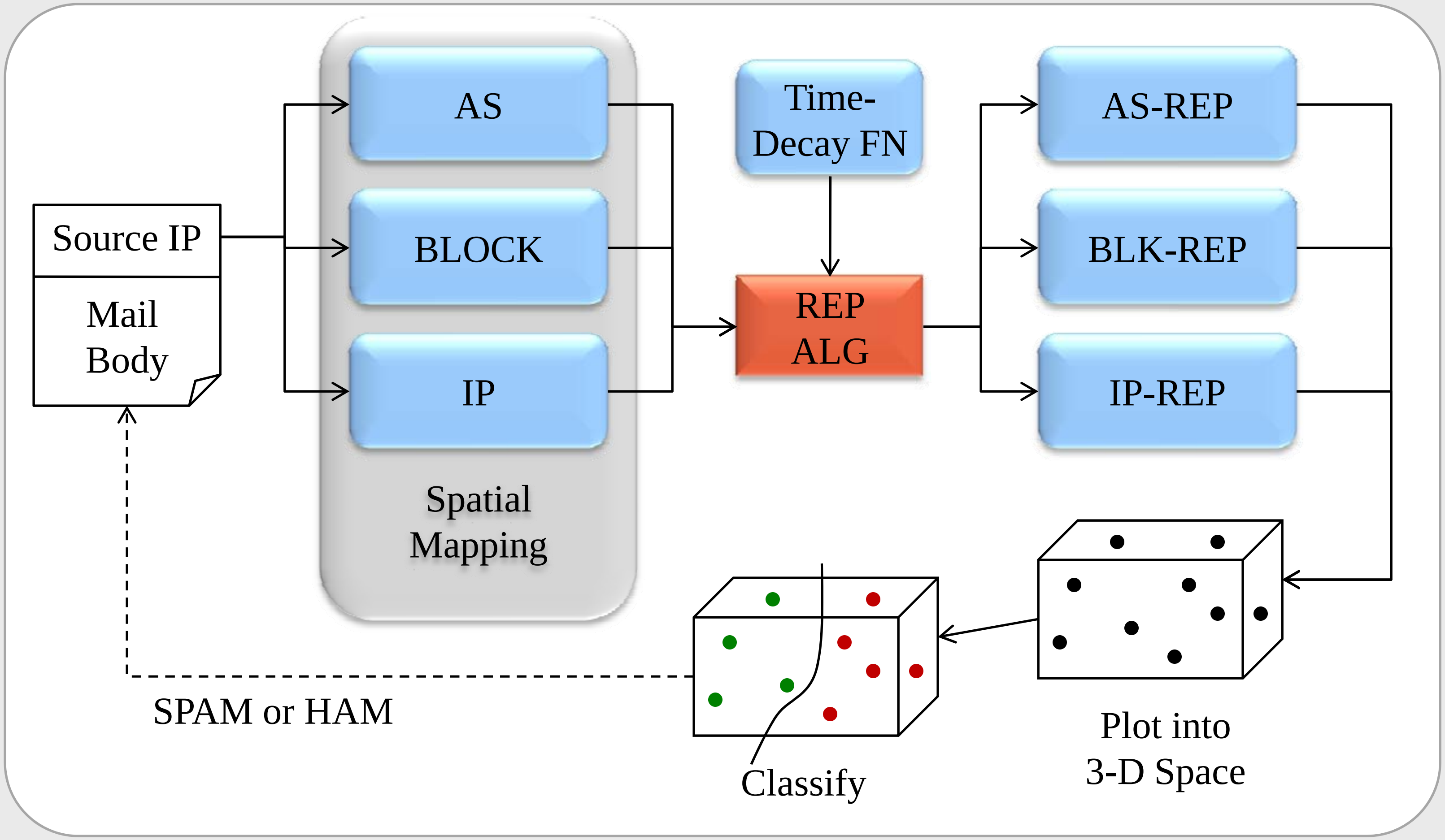


PreSTA: SPAM USAGE

PreSTA:
Preventative
Spatio-Temporal
Aggregation

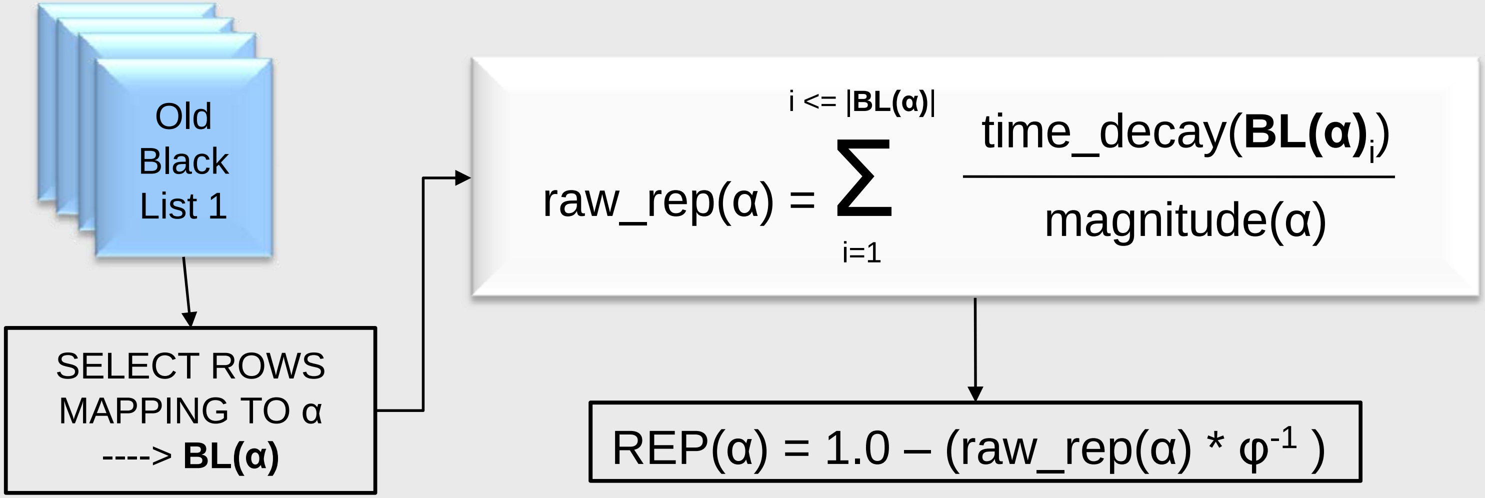


VALUATION WORKFLOW



REPUTATION ALGORITHM

- To calculate reputation for entity α:



Old Black List 1

SELECT ROWS MAPPING TO α ----> **BL(α)**

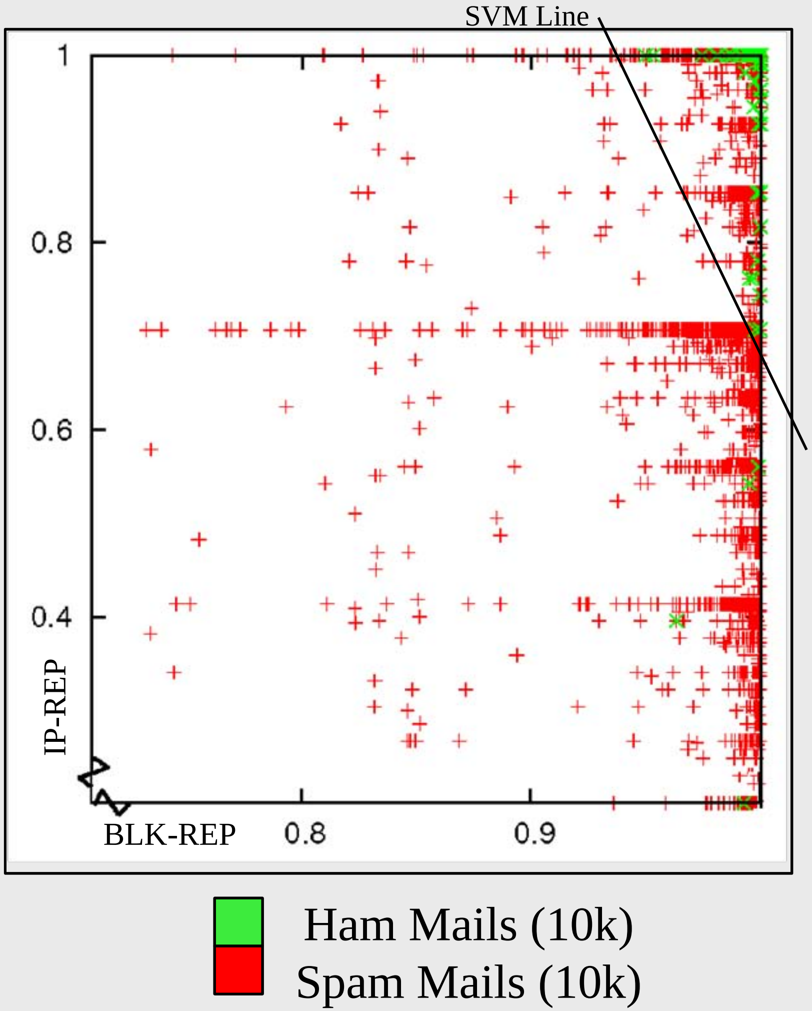
$$\text{raw\_rep}(\alpha) = \sum_{i=1}^{i <= |\mathbf{BL(\alpha)}|} \frac{\text{time\_decay}(\mathbf{BL(\alpha)}_i)}{\text{magnitude}(\alpha)}$$

$$\text{REP}(\alpha) = 1.0 - (\text{raw\_rep}(\alpha) * \varphi^{-1})$$

- time_decay(*): Returns on [0,1], higher weight to more recent events
- magnitude(α): Number of IPs in grouping α
- φ: Normalization constant putting REP() on [0,1]

SVM LEARNING

- Combination strategies
- Support Vector Machine
  - Supervised learning
  - Train over previous email to classify current emails
- Draws surface (threshold) best separating points
  - Can adjust penalty weight to keep false positives low
  - Polynomial, RBF kernels improve on linear performance



SPAM: TESTING DATASETS

BLACKLIST

- Subscribe to Spamhaus provider
- Process `diff`’s between lists into DB
- Scores 86.2% detection w/0.37% FP

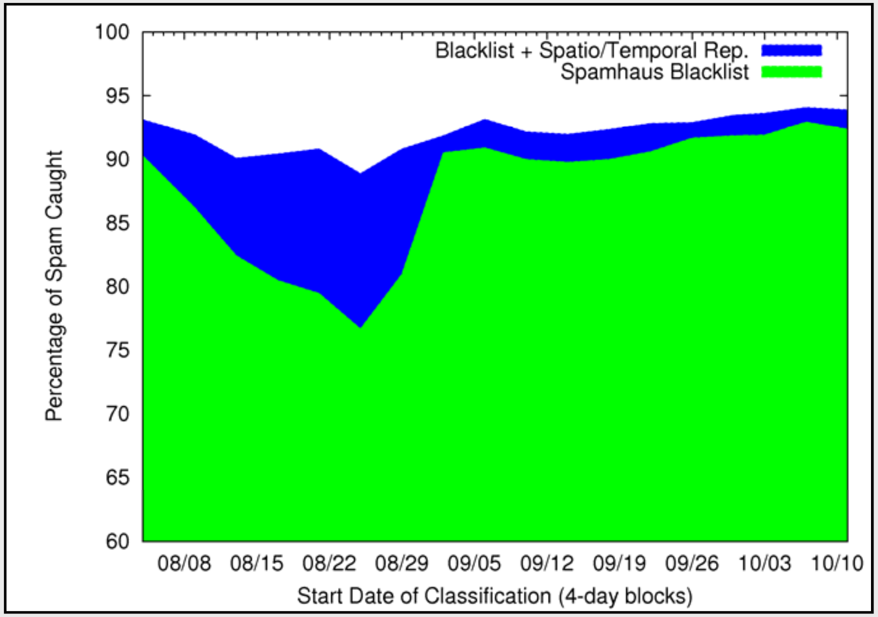
AS-MAP

- Use RouteViews data to map IP->AS

EMAIL

- 10 weeks: 15 mil. UPenn mail headers
- Proofpoint score as definitive spam/ham tag

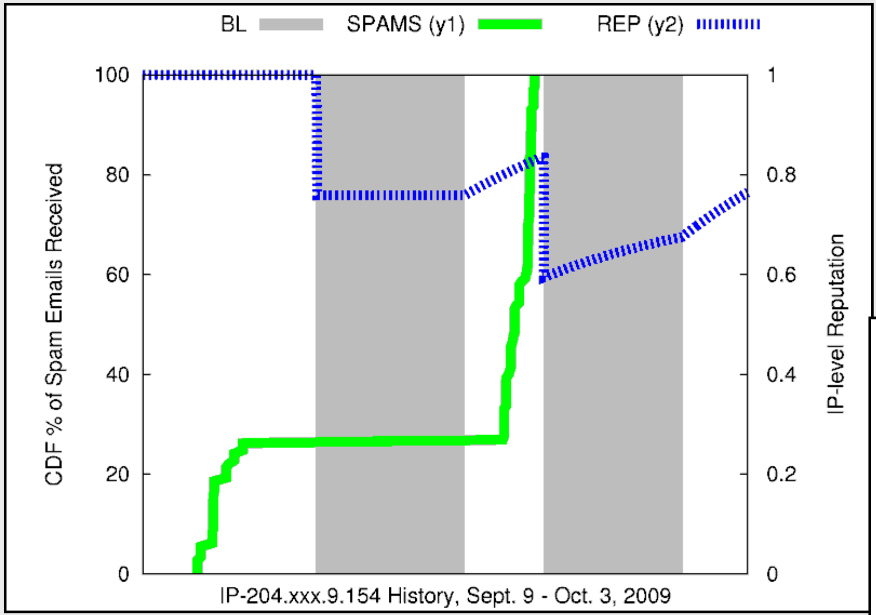
SPAM: PERFORMANCE (1)



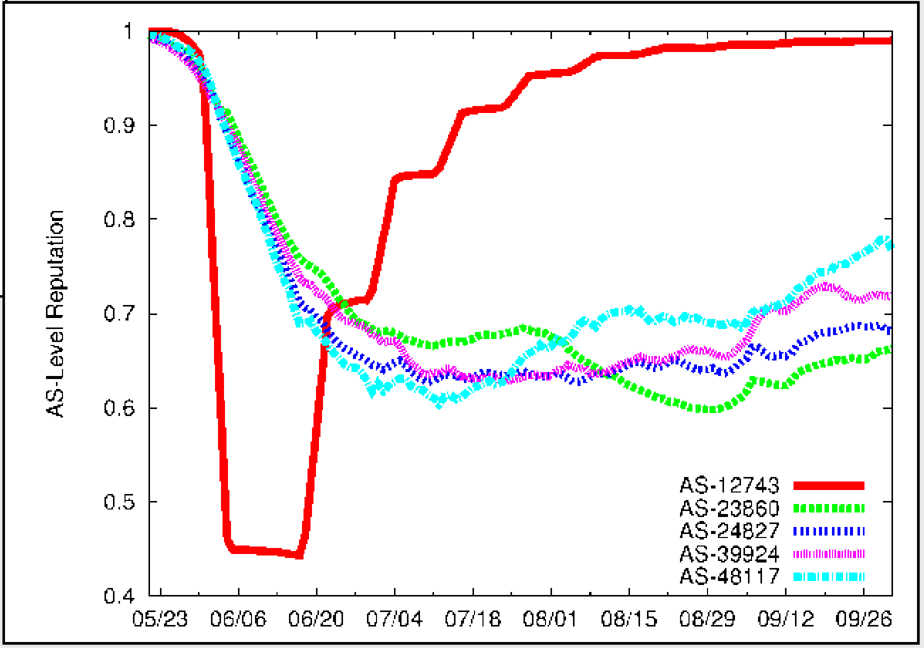
Captures up to 50% of mail not caught by traditional blacklists with the same low false-positives

- We capture between 20-50% of spam that gets past current blacklists
  - By design our FP-rate is equivalent to BLs: ~0.4%
- Total blockage remains near constant: 90%
  - Blacklists are reactive, we are predictive. We can cover its slack
  - Cat and mouse. Graph should roll over time

SPAM: PERFORMANCE (2)



< Temporal (single IP) example where our metric could mitigate spam reception



Probable botnet attack which our metric could mitigate via both temporal and spatial means >

SPAM: CONTRIBUTIONS

SNARE [3] (GA-Tech)

- Supervised learning across 13-network level features, including spatio-temporal ones
- Don't need blacklists (but neither do we, only known spamming IPs)

Existing ‘Reputation Systems’ [6]

- Exclusive use of negative feedback
- Existing email reputation systems [5] focus only on sharing classifications

DISTINGUISHING CONTRIBUTIONS

- Formalization of predictive spatio-temporal reputation
- Development of a lightweight mail filter, capable of 500k+ mails/hour

FUTURE: WIKIPEDIA

PURPOSE: Build a blacklist of user-names/IPs
based on the probability they will vandalize

TEMPORAL

- Straightforward, vandals are probably repeat offenders
- Registered users have IDs indicating when they joined, are new users more likely to vandalize?

SPATIAL

- Geographical: Based on user location (*i.e.*, Wash. D.C.)
- Topical: A user may vandalize one topic (Rush Limbaugh), while properly editing another (Barack Obama)
- Anonymous users: IP address properties

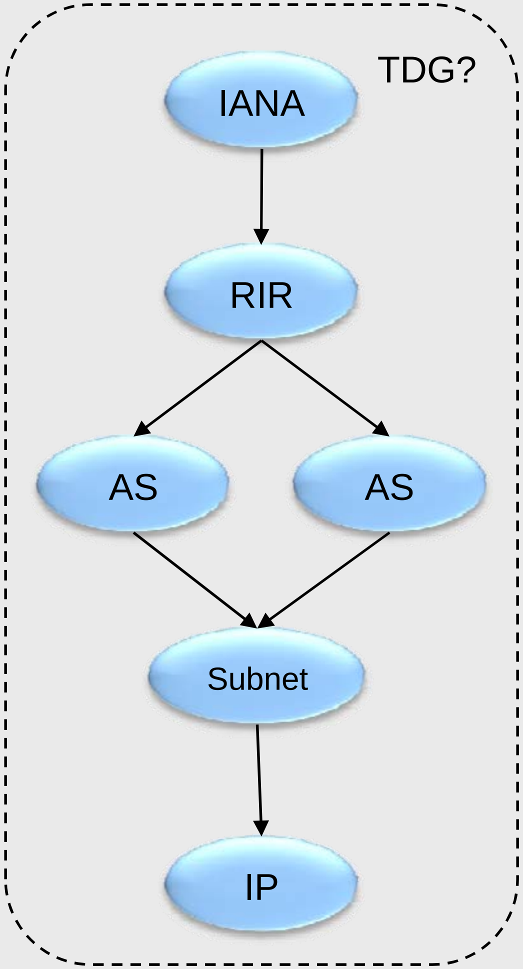
FEEDBACK

- Certain administrators have rollback (revert) privileges
- Comment: “Reverted edit by X to last edition by Y”

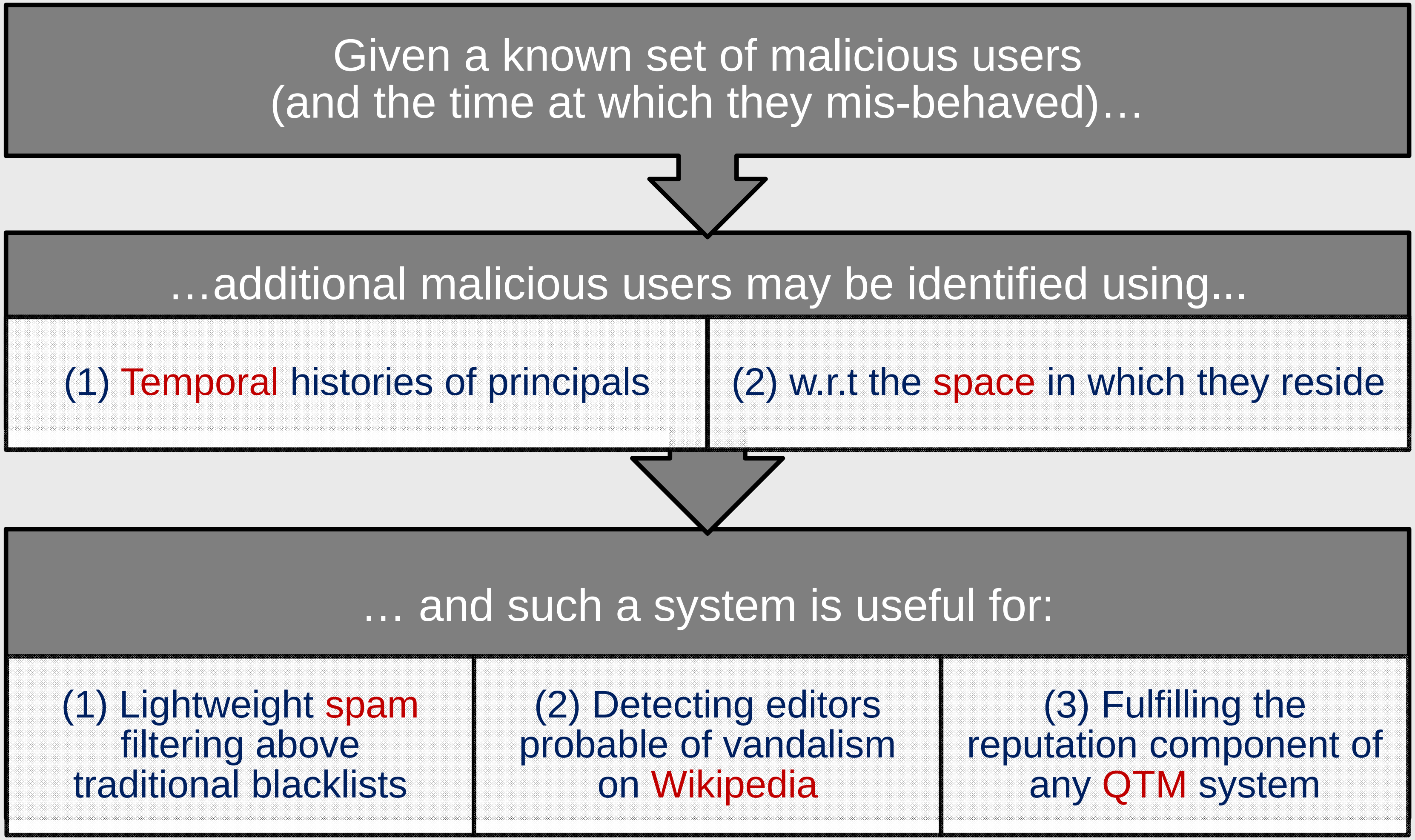
FUTURE: QUANTM [2] MODEL

- PreSTA may trivially fulfill the reputation component of qualifying QTM systems
  - TDG-like hierarchy of IP-delegation
  - Spatial groups from credential depth?

- General-use case criteria:
  - (1) There must be a grouping function to define finite sets of participants
  - (2) Observable and dynamic feedback sufficient to construct behavior history



CONCLUSIONS



Given a known set of malicious users
(and the time at which they mis-behaved)…

…additional malicious users may be identified using...

(1) Temporal histories of principals

(2) w.r.t the space in which they reside

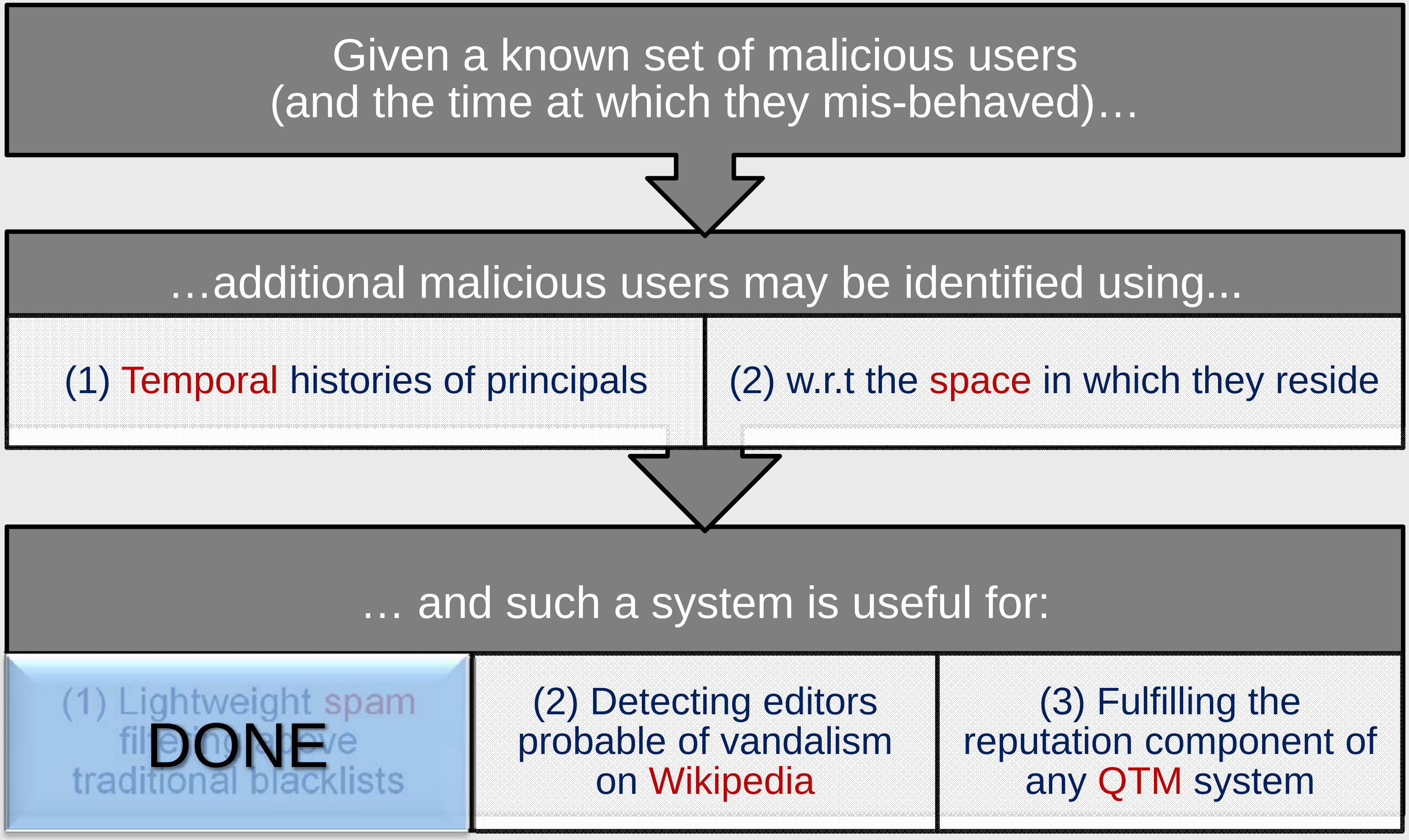
… and such a system is useful for:

(1) Lightweight spam filtering above traditional blacklists

(2) Detecting editors probable of vandalism on Wikipedia

(3) Fulfilling the reputation component of any QTM system

CONCLUSIONS



Given a known set of malicious users
(and the time at which they mis-behaved)…

…additional malicious users may be identified using...

(1) Temporal histories of principals

(2) w.r.t the space in which they reside

… and such a system is useful for:

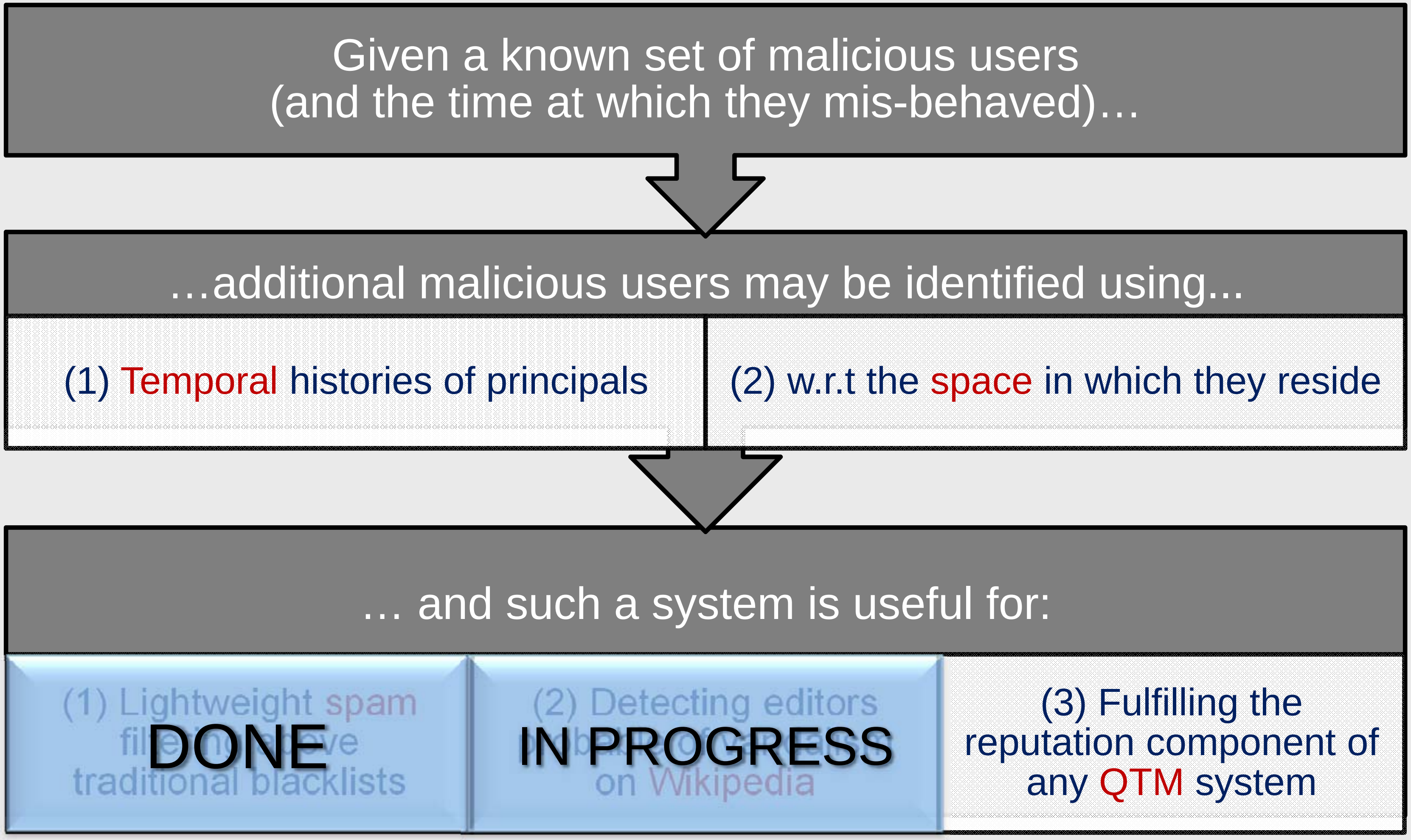
(1) Lightweight spam filtering above traditional blacklists

DONE

(2) Detecting editors probable of vandalism on Wikipedia

(3) Fulfilling the reputation component of any QTM system

CONCLUSIONS



Given a known set of malicious users
(and the time at which they mis-behaved)…

…additional malicious users may be identified using...

(1) Temporal histories of principals

(2) w.r.t the space in which they reside

… and such a system is useful for:

(1) Lightweight spam filtering to improve traditional blacklists

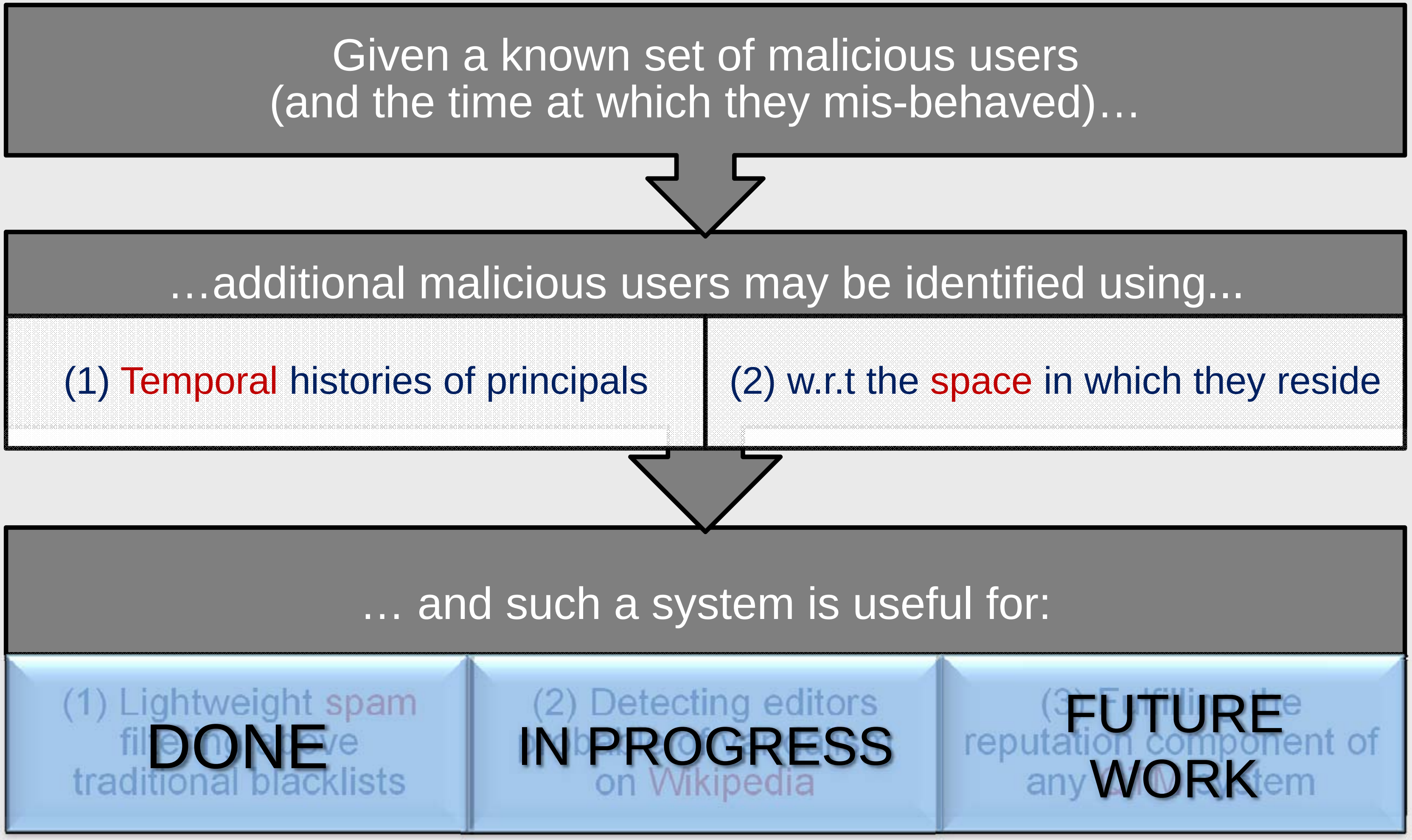
DONE

(2) Detecting editors [illegible] on Wikipedia

IN PROGRESS

(3) Fulfilling the reputation component of any QTM system

CONCLUSIONS



Given a known set of malicious users (and the time at which they mis-behaved)…

…additional malicious users may be identified using...

(1) Temporal histories of principals

(2) w.r.t the space in which they reside

… and such a system is useful for:

(1) Lightweight spam filtering above traditional blacklists

DONE

(2) Detecting editors [illegible] on Wikipedia

IN PROGRESS

(3) Fulfilling the reputation component of any [illegible] system

FUTURE WORK

REFERENCES

- [1] - West, A.G. *et al*. Preventing Malicious Behavior Using Spatio-Temporal Reputation. In submission to *EuroSys '10*.
- [2] - West, A.G. *et al*. QuanTM: A Quantitative Trust Management System. In Proceedings of *EuroSec '09*.
- [3] - Hao, S. *et al*. Detecting Spammers with SNARE: Spatio-temporal Network Level Automated Reputation Engine. In *18th USENIX Security Symposium*, August 2009.
- [4] - Ramachandran, A. *et al.* Understanding the Network-level Behavior of Spammers. In *SIGCOMM '06*.
- [5] - Alperovitch, D. *et al*. Taxonomy of Email Reputation Systems. In *Distributed Computing Systems Workshops '07*.
- [6] - Kamvar, S.D. et al. The EigenTrust Algorithm for Reputation Management in P2P Systems. In *12th WWW '03*.